An energy-efficient high-performance processor with reconfigurable data-paths using RSFQ circuits

Takagi, Naofumi

An energy-efficient high-performance processor with reconfigurable data-paths using RSFQ circuits

Naofumi Takagi
Graduate School of Informatics, Kyoto Univsersity
Yoshida-honmachi, Sakyo-ku, Kyoto, 606-8501, Japan
takagi@i.kyoto-u.ac.jp

Abstract

We show recent progress in our research on an energy-efficient high-performance processor with reconfigurable data-paths (RDPs) using rapid single-flux-quantum (RSFQ) circuits. We mainly describe the architectural details of an RDP implemented using RSFQ circuits. An RDP consists of a lot of floating-point units (FPUs) and operand routing networks (ORNs) which connect the FPUs. We reconfigure the RDP to fit a computation, i.e., a group of floating-point operations, appearing in a 'for' loop of programs for numerical computations by setting the route in ORNs before the execution of the loop. In the RDP, a lot of FPUs work in parallel with pipelined fashion, and hence, very high-performance computation is achieved.

Keywords: Rapid single-flux-quantum (RSFQ) circuit, High-performance computer, Reconfigurable data-path

1. Introduction

High-performance computers are indispensable for research and development in various fields, such as space science, aerospace engineering, materials science, molecular science, structural calculation, environment simulation, etc. In future high-performance computers, electrical power consumption will be one of the most serious problems. Therefore, development of an energy-efficient high-performance computer is desired.

Superconducting rapid single-flux-quantum (RSFQ) circuit technology [1] is expected to be a next generation circuit technology which enables ultra-high-speed computation with ultra-low power consumption [2]. It is attractive to develop a high-performance computer using RSFQ circuits. In order to make the most of the good properties of RSFQ circuits in a high-performance computer, we have to adopt computer architecture suitable for RSFQ implementation.

Recently, we proposed an energy-efficient high-performance processor with reconfigurable data-paths (RDPs) using RSFQ circuits [3]. An RDP consists of a lot of, e.g., a few thousand, floating-point units (FPUs) and operand routing networks (ORNs) which connect the FPUs [4]. We reconfigure an RDP to fit a computation, i.e., a group of floating-point operations, which appears in a 'for' loop in a program for numerical computation, by setting the route in ORNs before the execution of the loop. In an RDP, a lot of FPUs work in parallel with pipelined fashion, and hence, very high-performance computation is achieved. We have proposed to implement an RDP by RSFQ circuits. (We call it an SFQ-RDP.) Since data flow unidirectionally without feedback loops or conditional branches in it, an RDP is suitable for RSFQ implementation.

We have been developing basic technologies for an SFQ-RDP in a CREST-JST project [5]. We have investigated architectural details of an SFQ-RDP. We have developed a Nb 9-layer 1μm fabrication process with a device structure having the active layers at the top, two passive transmission line (PTL) layers in the middle, and the DC power layers at the bottom [6]. We have also developed a logic cell library for the new 1μm fabrication process [7,8], as well as CAD tools including a clock tree synthesizer and an automatic wire router [9,10]. We have designed and fabricated prototype SFQ LSIs of a floating-point adder and a multiplier, as well as simple RDPs, using SRL's conventional Nb 4-layer 2μm process and the newly developed Nb 9-layer 1μm process [11-15]. (SRL: Superconductivity Research Laboratory, International Superconductivity Technology Center, Japan.)

In this paper, we first present the outline of an SFQ-RDP, and then show the architectural details.

2. Outline of an SFQ-RDP

RSFQ circuits have good properties of high-speed switching, high-speed signal transmission, low power consumption and hence low heat radiation, etc. Since an SFQ pulse is used as the carrier of information, RSFQ digital circuits work by pulse logic. Therefore, each logic gate of RSFQ digital circuits is a clocked gate and has a function of latch. RSFQ digital circuits are suitable for pipeline processing on streaming data, but are not suitable for processing with feedback loops and/or conditional branches. In order to make the most of the good properties of RSFQ circuits in a high-performance computer, we have to adopt computer architecture suitable for RSFQ implementation.

In the development of a high-performance computer, the memory-wall problem is one of the greatest obstacles [16,17]. The memory-wall problem is the problem that the memory bandwidth cannot be wide enough related to the processor performance because of the gap between the operating speed of a processor and that of a memory, and hence, the performance of a computer is limited. In an RSFQ computer, this problem would be more crucial, because RSFQ digital circuits operate very fast while a large-scale superconductive random access memory seems difficult to be implemented.

In general, higher memory bandwidth is required when the capacity of arithmetic operations increases. A large amount of load/store operations concerns the read/write operations on the intermediate data. Therefore, we can reduce the required memory bandwidth by eliminating the load/store operations corresponding to the read/write operations on the intermediate data. In order to realize this, we proposed a high-performance processor with reconfigurable data-paths (RDPs).

A reconfigurable data-path (RDP) consists of a lot of floating-point units (FPUs) and operand routing networks (ORNs) which connect the FPUs [3], as shown in Fig. 1. We attach RDPs to a general purpose processor as an accelerator as shown in Fig. 2. Data are fed to an RDP from the memory via streaming buffers (SBs). We reconfigure an RDP to fit a computation, i.e., a group of floating-point operations, which appears in a ‘for’ loop in a program for numerical computation, by setting the route in ORNs before the execution of the loop. Namely, a compiler extracts a data-flow graph from each computation-intensive loop body in a program, and produces configuration data of the RDP based on the graph. Then, in execution of the compiled program which includes several loops, the RDP is dynamically reconfigured according to the configuration data just before entering each loop. Since a result of an FPU is forwarded to another FPU in the next row through an ORN, we do not need to store/load the intermediate result to/from memory. Furthermore, we do not need to fetch the arithmetic instructions from memory. Therefore, we can reduce the required memory bandwidth drastically. In an

RDP, a lot of FPUs operate in parallel with pipelined fashion, and hence, very high performance computation is achieved.

We have proposed to implement the RDPs and the SBs by RSFQ circuits. We call an RDP implemented by RSFQ circuits as an SFQ-RDP. Since data flow unidirectionally without feedback loops nor conditional branches in it, an RDP is suitable for RSFQ implementation. By implementing RDPs with RSFQ circuits, we can develop an energy-efficient high-performance computer.

3. Architectural details of an SFQ-RDP

We have determined architectural details of an RDP, by analyzing various programs for numerical computations [18].

We provide two types of FPU. One is a floating-point adder (FPA) unit, and the other is a floating-point multiplier (FPM) unit. Each FPU has three inputs, A, B, and C, and produces three outputs (A*B or A), B, and C, where the operation * is addition in an FPA unit and multiplication in an FPM unit. We arrange FPA units and FPM units alternately in a row.

We have classified the programs into three categories with respect to the size of the largest loop body in them, and have determined three types of RDP correspondingly, as shown in Table 1. Here, 'Width' is the number of FPUs in a row, 'Height' is the number of rows, and 'MCL (maximum connection length)' is the maximum horizontal distance between an FPU and a one in the next row connected to it through an ORN.

Table 1: Three types of reconfigurable data-path

Type	# Input	# Output	Width	Height	MCL
RDP-S	19	12	22	14	4
RDP-M	19	12	24	17	5
RDP-L	38	24	41	34	6

We have found that the bit-serial architecture is suitable for an SFQ-RDP because of the properties of RSFQ circuits [19]. We have adopted the bit-serial architecture. We employ the data format shown in Fig. 3 for representing an IEEE754 floating-point number in an SFQ-RDP. We use two bit-sequences; one for the sign and the exponent and the other for the significand. In the former, following the sign bit (S), the exponent (E) appears in lsb-first fashion. In the latter, the significand (F) appears in lsb-first fashion. The hidden-1 is represented explicitly. For single-precision, E is of 8-bits and F with the hidden-1 is of 24-bits. Therefore, an operand is of 24-bits. For double-precision,

E is of 11-bits and F with the hidden-1 is of 53-bits. Therefore, an operand is of 53-bits.

A floating-point adder (FPA) consists of a sign&exponent processing part and a significand processing part [11]. The latter consists of an adder/subtracter which is basically a bit-serial adder, and slightly complicated shifters for alignment and for normalization. A floating-point multiplier (FPM) also consists of a sign&exponent processing part and a significand processing part [12]. The main component of the latter is a bit-serial multiplier based on a systolic algorithm. The latency, i.e., the number of clock cycles between the first bit input and the first bit output, of the FPA, as well as that of the FPM, is $2n+1$, where n is the length of the operand (F with hidden-1). The minimum input interval, i.e., the number of clock cycles between the first bit input and the first bit input of the next operand, is $n+2$. Note that it takes n clock cycles to input or output an operand.

We have designed and fabricated a half-precision floating-point adder and a multiplier, where E is of 5-bits and F with the hidden-1 is of 11-bits, using SRL's conventional Nb 4-layer 2μm fabrication process. The former consists of 10,223 Josephson junctions, is of 5.86mm x 5.72mm in size, and has operated correctly at 20GHz for typical test inputs [11]. The latter consists of 11,044 Josephson junctions, is of 6.22mm x 3.78mm in size, and has operated correctly at 31.5GHz for typical test inputs [12]. We are now designing a floating-point adder and a multiplier which are to be fabricated using the new Nb 9-layer 1μm fabrication process, targeting 50GHz operation. Several component circuits have been fabricated, which operate at about 80GHz [13].

There are two major candidates of implementation of an operand routing network (ORN); one is a set of multiplexors and the other is a network of switches. We adopt the latter because it is superior in scalability and in timing adjustment. We implement an ORN by arranging 2x2 switches (CBs) in the checker pattern as shown in Fig. 4 [20]. Each ORN consists of 4x'MCL'+1 rows of 3x'Width' switches. Each switch realizes one of the four functions, (A, B) to (A, B) or (B, A) or (A, A) or (B, B) according to the configuration data. Each CB has two D flip-flops (D-FFs) for storing the corresponding configuration data. The D-FFs of CBs in each row of an ORN are connected in cascade, and configuration data for them are fed bit-serially.

We have designed and fabricated a 2x3 RDP with six 7-bit arithmetic and logic units (ALUs) instead of FPUs using SRL's conventional Nb 4-layer 2μm fabrication process. It consists of 14,040 Josephson junctions, is of 6.84mm x 6.72mm in size and has operated at 23GHz [14]. All ALUs have performed all ALU functions except subtraction correctly for typical test inputs, and the ORNs have worked correctly for all

routes. We have also designed and fabricated a 2x2 RDP with four ALUs using the new Nb 9-layer 1μm fabrication process. It consists of 11,458 Josephson junctions, is of 5.61mm x 2.82mm in size, and has operated at 45GHz [15]. All ALUs and ORNs have worked correctly. We are now designing a 4x4 RDP with 16 ALUs targeting 50GHz operation.

4. Conclusion

Recently, we proposed an energy-efficient high-performance processor with reconfigurable data-paths (RDPs) using RSFQ circuits and have been developing basic technologies for it in a CREST-JST project. We have investigated architectural details of an SFQ-RDP, have developed a Nb 9-layer 1μm fabrication process and a logic cell library for it, as well as CAD tools. We have also designed and fabricated several prototype SFQ LSIs. Through the development, we have shown feasibility and effectiveness of an SFQ-RDP.

According to a rough estimation based on our studies, using a future 0.5μm fabrication process, we will be able to integrate one row of 32 double-precision FPUs (16 FPA units and 16 FPM units) and one ORN, which consist of about 1.8MJJs in total and operate at 100GHz with consuming electric power of about 50mW, on a chip. We will be able to mount 32 SFQ-RDP chips and two SB chips on a multi-chip module (MCM) which provides superconducting chip-to-chip interconnections. Consequently, one MCM will contain an RDP including 1024 FPUs, i.e., a square array of 32 by 32 FPUs. The total peak performance and power consumption of the MCM are estimated to be about 2TFLOPS and 1.6W, respectively.

Acknowledgements

The author would like to express his sincere gratitude to Prof. Akira Fujimaki of Nagoya University, Prof. Nobuyuki Yoshikawa of Yokohama National University, Prof. Kazuaki Murakami of Kyushu University, Associate Prof. Kazuyoshi Takagi of Kyoto University, and all other members of the CREST-JST project 'Low-power, high-performance, reconfigurable processor using single-flux-quantum circuits.' This work is supported by CREST-JST.

References

[1] K. K. Likharev, V. K. Semenov, RSFQ logic/memory family: A new Josephson-junction technology for sub-terahertz-clock-frequency digital systems, IEEE Trans. Appl. Supercond., 1, 1 (1991) 3-28.

[2] N. Yoshikawa, Recent development and perspective of ultra-high-speed microprocessors using single-flux-quantum circuits. IEICE Trans. Electronics, J91-C, 3 (2008) 183-193 (in Japanese).

[3] N. Takagi, K. Murakami, A. Fujimaki, N. Yoshikawa, K. Inoue, H. Honda, Proposal of a desk-side supercomputer with reconfigurable data-paths using rapid single-flux-quantum circuits, IEICE Trans. Electronics, E91-C, 3 (2008) 350-355.

[4] K. Shimasaki, T. Nagano, H. Honda, F. Mehdipour, K. Inoue, K. Murakami, On-chip network architecture for large scale reconfigurable datapath, IPSJ SIG Technical Reports, 2007-ARC-173 (2007) 115-120 (in Japanese).

[5] http://www.jst.go.jp/kisoken/crest/en/area04/5-06.html

[6] S. Nagasawa, T. Satoh, K. Hinode, Y. Kitagawa, M. Hidaka, H. Akaike, A. Fujimaki, K. Takagi, N. Takagi, N. Yoshikawa, New Nb multi-layer fabrication process for large-scale SFQ circuits, Physica C, 469 (2009) 1578-1584.

[7] H. Akaike, M. Tanaka, K. Takagi, I. Kataeva, R. Kasagi, A. Fujimaki, K. Takagi, M. Igarashi, H. Park, Y. Yamanashi, N. Yoshikawa, K. Fujiwara, S. Nagasawa, M. Hidaka, N. Takagi, Design of Single Flux Quantum cells for a 10-Nb-layer process, Physica C, 469 (2009) 1670-1673.

[8] Y. Yamanashi, T. Kainuma, N. Yoshikawa, I. Kataeva, H. Akaike, A. Fujimaki, M. Tanaka, N. Takagi, S. Nagasawa, M. Hidaka, 100GHz demonstrations based on the single-flux-quantum cell library for the 10 kA/cm^2 Nb multi-layer process, IEICE Trans. Electronics., E93-C, 4 (2010) 440-444.

[9] K. Takagi, Y. Ito, S. Takeshima, M. Tanaka, N. Takagi, Layout-driven skewed clock tree synthesis for superconducting SFQ circuits, IEICE Trans. Electronics, E94-C, 3 (2011), 288-295.

[10] M. Tanaka, K. Obata, Y. Ito, S. Takeshima, M. Sato, K. Takagi, N. Takagi, H. Akaike, A. Fujimaki, Automated passive-transmission-line routing tool for single-flux- quantum circuits based on A* algorithm, IEICE Trans. Electronics., E93-C, 4 (2010) 435-439.

[11] H. Park, Y. Yamanashi, K. Taketomi, N. Yoshikawa, M. Tanaka, K. Obata, Y. Ito, A. Fujimaki, N. Takagi, K. Takagi, and S. Nagasawa, Design and implementation and on-chip high-speed test of SFQ half-precision floating-point adders, IEEE Trans. Appl. Supercond., 19, 3 (2009) 634-639.

[12] H. Hara, et al., Design, implementation and on-chip high-speed test of SFQ half-precision floating-point multiplier, IEEE Trans. Appl. Supercond., 19, 3 (2009), 657-660.

[13] T. Kainuma, Y. Shimamura, F. Miyaoka, Y. Yamanashi, N. Yoshikawa, A. Fujimaki,

K. Takagi, N. Takagi, S. Nagasawa, Design and implementation of component circuits of an SFQ half-precision floating-point adder using 10 kA/cm^2 Nb process, IEEE Trans. Appl. Supercond., 21, 3 (2011), 827-830.

[14] A. Fujimaki, R. Kasagi, K. Takagi, I. Kataeva, H. Akaike, M. Tanaka, N. Takagi, N. Yoshikawa, K. Murakami, Demonstration of 2x3 reconfigurable-data-path processors with 14000 Josephson Junctions, Proc. 12th International Superconductive Electronics Conference (ISEC '09) (2009) SP-O4.

[15] M. Okada, I. Kataeva, M. Ito, M. Tanaka, H. Akaike, A. Fujimaki, N. Yoshikawa, S. Nagasawa, N. Takagi., Demonstration of a 2x2 single-flux-quantum reconfigurable data-path based on the 10-kA/cm^2 process, IEICE Technical Report,SCE2010-33, (2010) (in Japanese).

[16] W. A. Wulf, S. A. McKee, Hitting the memory wall: implications of the obvious, ACM SIGARCH Computer Architecture News, 23, 1 (1995) 20-24.

[17] D. Burger, J. R. Goodman, A. Kagi, Memory bandwidth limitations of future micro-processors, Proc. 23rd Annual International Symposium on Computer Architecture (1996) 78-89.

[18] F. Mehdipour, H. Honda, K. Inoue, H. Kataoka, K. Murakami, A design scheme for a reconfigurable accelerator implemented by single-flux-quantum circuits, J. Systems Architecture - Embedded Systems Design 57 (2011) 169-179.

[19] N. Takagi, M. Tanaka, Comparisons of synchronous-clocking SFQ adders, IEICE Trans. Electronics. E93-C, 4 (2010) 429-434.

[20] I. Kataeva, H. Akaike, A. Fujimaki, N. Yoshikawa, N. Takagi, K. Inoue, H. Honda, K. Murakami, An operand routing network for an SFQ reconfigurable data-paths processor, IEEE Trans. Appl. Supercond., 19, 3 (2009) 665-669.

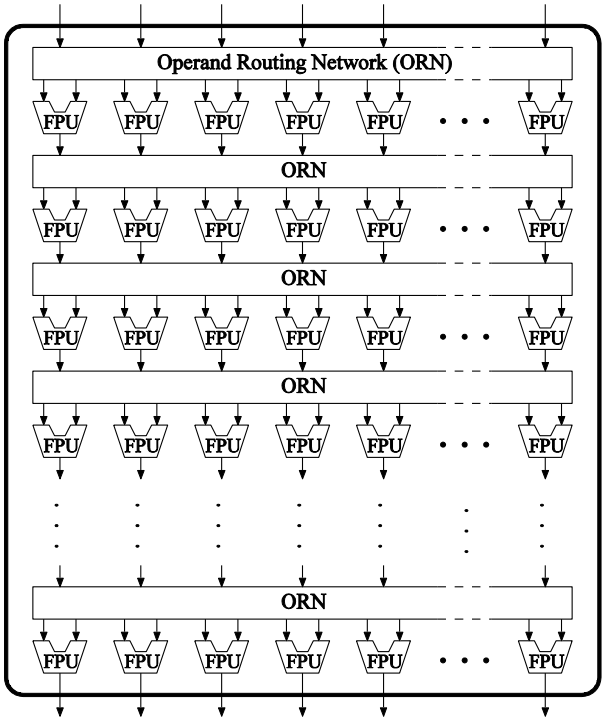

Figure 1: A reconfigurable data-path (RDP)

An RDP consists of a lot of floating-point units (FPUs) and operand routing networks (ORNs) which connect the FPUs.

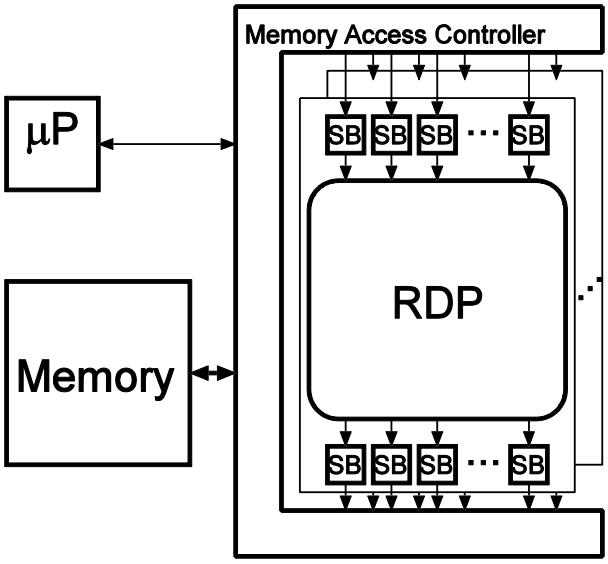

Figure 2: A processor with RDPs

RDPs are attached to a general purpose processor as an accelerator.

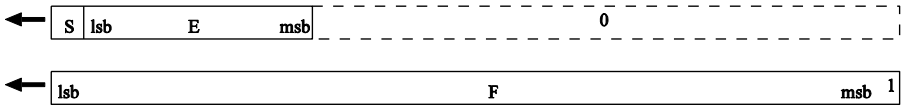

Figure 3: Data format of a floating-point number

Two bit-sequences are used for representing an IEEE floating-point number. In one, following the sign bit (S), the exponent (E) appears in lsb-first fashion. In the other, the significand (F) appears in lsb-first fashion. The hidden-1 is represented explicitly.

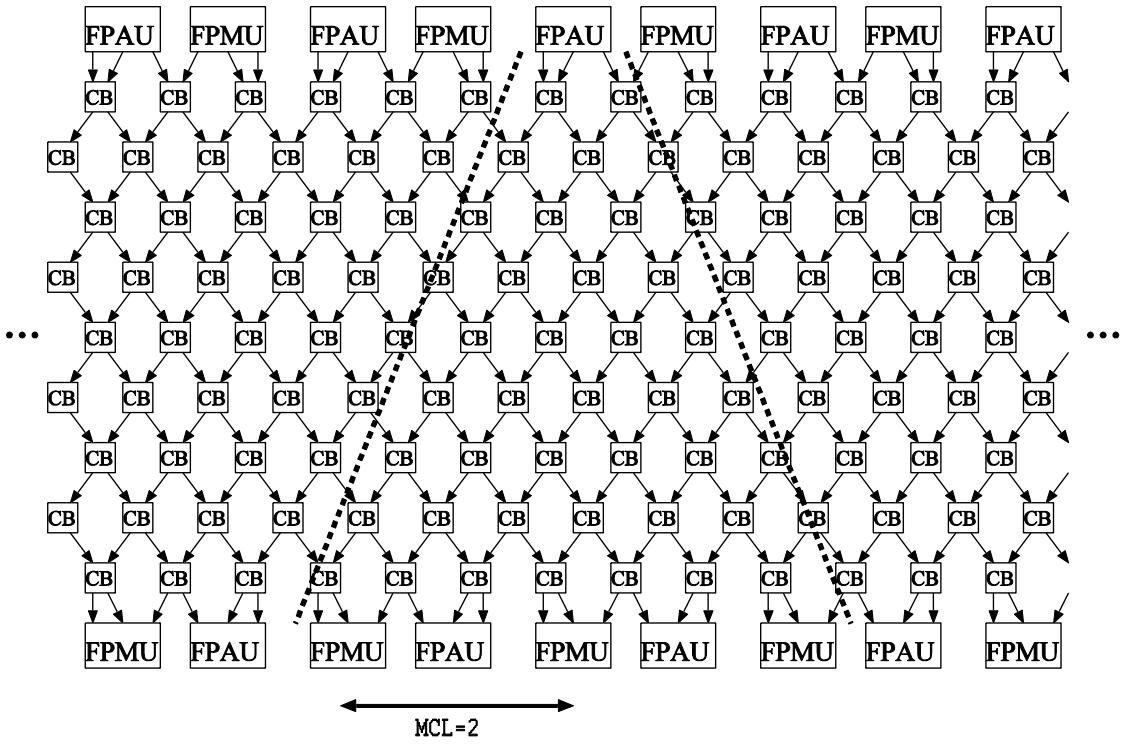

Figure 4: A part of an operand routing network (ORN) with MCL=2

An ORN is implemented as a network of CBs (2x2 switches). CBs are arranged in the checker pattern.